Building Viewpoints in an Object-Based Representation System for Knowledge Discovery in Databases

Arnaud Simon and Amedeo Napoli
LORIA-UMR 7503
615, rue du Jardin Botanique,
BP 101, 54600 Villers-lès-Nancy, France
{*a-simon,napoli*}*@loria.fr*

Abstract

In this paper, we present an approach to knowledge discovery in databases in the context of object-based representation systems. The goal of this approach is to extract viewpoints and association rules from data represented by objects. A viewpoint is a hierarchy of classes and an association rule can be defined within a viewpoint or between two classes lying in different viewpoints. The viewpoints construction algorithm allows to manipulate objects which are indifferently classes or instances of classes. Moreover, multivalued attributes as well as relations can be taken into account in the current approach.
Keywords: KDD, *Galois lattice, object-based representation system, association rules.*

1 Introduction

Knowledge discovery in databases (KDD) aims at the discovery of useful information from large collections of data [3]. The discovered knowledge is represented by classification, association or causal rules describing properties of the data, classes or clustering of the objects in the database, etc. A KDD system relies on five main components: (1) data (and database management system); (2) *data mining* modules for extracting knowledge; (3) visualisation modules showing both the data and the knowledge; (4) a knowledge-based system able to solve problems on the domain of data; and (5) the human analyst who controls the process of knowledge discovery.

In the following, we present a KDD process whose goal is to discover clusterings and association rules from a collection of medical data. Currently, we are working on a medical database: the childhood cancer registry of Lorraine (a region of France). This database memorises descriptions of childhood cancer. The KDD process is then organised around an analyst (actually a physician) and a knowledge-based system describing the medical universe of the data being analysed. Actually, knowledge is represented in an *object-based representation system* (OBRS) where individual objects are clustered into classes organised in an inheritance hierarchy. Moreover, a classification process, inspired from description logics [2] can be used to draw inferences and solve problems in the medical universe.

In our approach, rough data are transformed into individual objects and then clustered into classes in order to be handled by the KDD process. Classes are then organised into special hierarchies that can be considered as special viewpoints on rough data. Association rules can be extracted from those hierarchies, rules inside one hierarchy and rules between hierarchies. Viewpoints and rules are proposed to the analyst for further validation.

The present work relies on several previous work on Galois lattices and rules extraction [8, 4, 6, 5]. Our approach extend these previous works in several directions. We work with an OBRS and we handle multivalued attributes and any type of relations (not only boolean relations or monovalued attributes). For a KDD perspective, our algorithm has to deal with a large amount of data. Nonetheless, Galois lattice construction is time and space expensive with respect to the number of handling data. Then, viewpoints are not only constructed from basic objects but also from classes (clusters of basic objects) which already summarise a large set of data. Lastly, we take advantage of domain knowledge during the Galois lattice construction. Domain knowledge allows us to obtain more precise and more fine-grained clustering of objects.

The remainder of this paper is organised as follows. First, we present the basis of object-based representation systems. Then, we introduce the notion of Galois viewpoints and the Galois viewpoints construction

process. The next section describes the association rule extraction process. A conclusion end the paper.

2 Object-Based Representation

In order to exploit domain knowledge, a KDD system has to be associated with a knowledge-based representation system (as explained in [3]). In our approach, we rely on an object-based representation system OBRS. Domain knowledge is represented by a *conceptual hierarchy* $\mathcal{H} = (\mathcal{X},\preccurlyeq,\omega)$ where $\mathcal{X}$ is a set of *classes*, $\preccurlyeq$ is a partial ordering — also called *subsumption* — and ω is the root of $\mathcal{H}$.

A class $\alpha \in \mathcal{X}$ represents a real-world concept and is defined by a name and a set of properties describing the behavioral and definitional characteristics of the represented concept. A class is used as a cluster for a set of *instances* or individuals, i.e. basic level objects. From a KDD perspective, an OBRS can be considered at two levels: the *intensional* level where classes represent real-world concepts on a given domain, and the *extensional* level where individuals represent data to be analysed.

The extension of a class α, denoted by $\texttt{Ext}(\alpha)$, is the set of instances of α. An instance $\texttt{i}$ of a class α verifies all the properties of the intension of α. The intension of a class α, denoted by $\texttt{Int}(\alpha)$, is defined by a set of attributes, denoted by $\texttt{A}(\alpha)$, which corresponds to the set of constraints the instances of α have to verify. Attributes can be *symbolic* attributes and filled with elementary values or *relations* establishing a correspondence between two classes. A range $\texttt{range}(\texttt{a},\alpha)$, describing the set of possible values for an attribute $\texttt{a}$ in the class α, is associated with $\texttt{a}$.

When an attribute $\texttt{a}$ is a relation, $\texttt{range}(\texttt{a},\alpha)$ denotes the class with which $\texttt{a}$ establishes a relation. For example, let us consider the conceptual hierarchy $\mathcal{H}1 = (\mathcal{X}1,\ \preccurlyeq,\ \omega 1)$ given in figure 1. The attribute `profession` in the class `Person` establishes a relation between `Person` and `Activity`.

In the range of an attribute $\texttt{a}$, certain values must be necessarily associated with $\texttt{a}$. The special keyword `required` introduce compulsory values and $\texttt{required}(\texttt{a},\alpha)$ denotes the set of compulsory values for the attribute $\texttt{a}$ in the class α. For example, $\texttt{required(diploma,Graduate)} = \{\texttt{MA},\texttt{MS}\}$.

The subsumption relation is defined on the set $\mathcal{X}$ of classes and can be considered at the extensional and intensional levels. At the extensional level, a class β subsumes a class α, denoted by $\alpha \preccurlyeq \beta$, if and only if an instance of α is also an instance of β, what means: $\texttt{Ext}(\alpha) \subseteq \texttt{Ext}(\beta)$. Then, α is described by at least the attributes of β and an attribute can have more possible values for β than for α. what means $\texttt{Int}(\alpha) \sqsupseteq \texttt{Int}(\beta)$.

3 Construction of a Viewpoint

In our approach, domain knowledge is represented by a conceptual hierarchy $\mathcal{H} = (\mathcal{X},\preccurlyeq,\omega)$. Basic data are reified as basic objects. A way of discovering new knowledge units in given sets of properties $\mathcal{A}$ and objects $\mathcal{D}$ (classes or instances) is to build a conceptual hierarchy, called a *viewpoint* and denoted by $\mathcal{H}_{\mathcal{D},\mathcal{A}} = (\mathcal{X}_{\mathcal{D},\mathcal{A}}, \preccurlyeq, \omega_{\mathcal{D},\mathcal{A}})$.

The construction of viewpoints relies on the Galois lattice theory. A lattice is a partially ordered set where each pair of elements $(\texttt{a},\texttt{b})$ has unique meet $(\texttt{a} \wedge \texttt{b})$ and join $(\texttt{a} \vee \texttt{b})$ elements [8]. Given a set of objects $\mathcal{D}$ and a set of attributes $\mathcal{A}$, the viewpoint extraction process builds a conceptual hierarchy $\mathcal{H}_{\mathcal{D},\mathcal{A}}$, where a class α is defined as a pair $(\texttt{Ext}(\alpha),\ \texttt{Int}(\alpha))$. The intension $\texttt{Int}(\alpha)$ is the most general intension w.r.t. $\sqsupseteq$ describing properties common to all objects laying in $\texttt{Ext}(\alpha)$ and $\texttt{Ext}(\alpha)$ is the set of all objects in $\mathcal{D}$ which have an intension more general than $\texttt{Int}(\alpha)$ w.r.t. $\sqsupseteq$. Since objects of $\mathcal{D}$ can be classes, it is the extension of the objects of $\mathcal{D}$ which are considered in $\texttt{Ext}(\alpha)$. An instance $\texttt{i}$ is then considered like a special class which verifies $\texttt{Ext}(\texttt{i}) = \{\texttt{i}\}$.

From a computational perspective, the algorithm used to build viewpoints is based on the GALOIS algorithm described in [4]. Given a set of objects $\mathcal{D}$ and a set of attributes $\mathcal{A}$, the algorithm tries to build a partial Galois lattice. The symbolic attributes and the relations are manipulated in a special way as described below. The algorithm proceeds in an incremental way and updates the current lattice as soon as new objects are added. When a new object $\texttt{o}$ is added, every class α of the hierarchy is examined. If $\texttt{Int}(\texttt{o})$ is more general than $\texttt{Int}(\alpha)$ then $\texttt{o}$ is added to $\texttt{Ext}(\alpha)$, else a new class β is created; given the set $\texttt{Ext}(\alpha) \cup \texttt{Ext}(\texttt{o})$, β is the class with the most general intension w.r.t. $\sqsupseteq$. A classification process is then used to introduce the new class β in the hierarchy [2].

3.1 The Manipulation of Attributes

Given a set $\mathcal{D}$ of objects and a set $\mathcal{A}$ of attributes, the class with the most general intension, denoted by $\texttt{MGI}_{\mathcal{A},\mathcal{D}}$, is defined as follows. The set of attributes of $\texttt{MGI}_{\mathcal{A},\mathcal{D}}$ is : $\texttt{A}(\texttt{MGI}_{\mathcal{A},\mathcal{D}}) = \bigcap_{\texttt{o}\in\mathcal{D}} \texttt{A}(\texttt{o}) \cap \texttt{A}$. The constraints on attributes are: $\texttt{range}(\texttt{a},\texttt{MGI}_{\mathcal{A},\mathcal{D}}) = \bigcup_{\texttt{o}\in\mathcal{D}} \texttt{range}(\texttt{a},\texttt{o})$

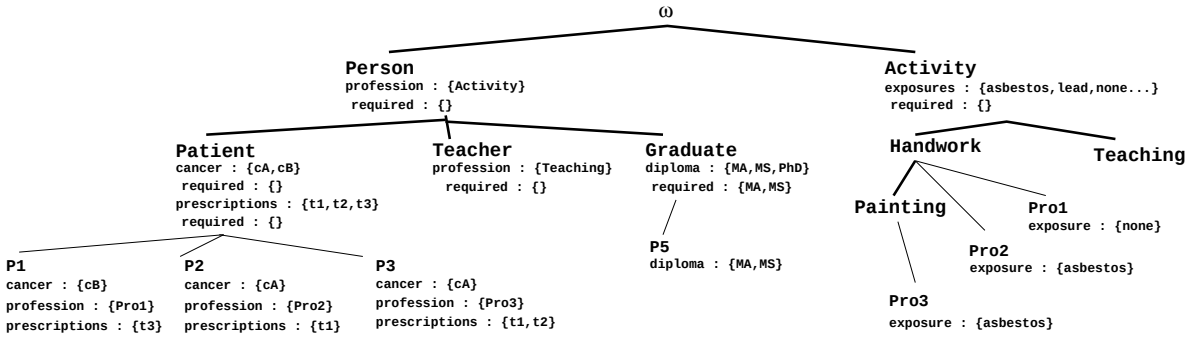

Figure 1: The hierarchy $\mathcal{H}1$ of `Person` and `Activity`.

and $\texttt{required}(\texttt{a}, \texttt{MGI}_{\mathcal{A},\mathcal{D}}) = \bigcap_{\texttt{o} \in \mathcal{D}} \texttt{required}(\texttt{a}, \texttt{o})$. For example, let us assume that $\mathcal{A}$ = {`cancer,prescriptions`} and $\mathcal{D}$ = {`P2,P3`}. The class with the most general intension $\texttt{MGI}_{\mathcal{A},\mathcal{D}}$ associated with $\mathcal{A}$ and $\mathcal{D}$ is defined by: $\texttt{range(cancer,MGI}_{\mathcal{A},\mathcal{D}}) = \{\texttt{cA}\}$, $\texttt{required(cancer,MGI}_{\mathcal{A},\mathcal{D}}) = \{\texttt{cA}\}$, $\texttt{range(prescriptions,MGI}_{\mathcal{A},\mathcal{D}}) = \{\texttt{t1,t2}\}$, $\texttt{required(prescriptions, MGI}_{\mathcal{A},\mathcal{D}}) = \{\texttt{t1}\}$. Taking into account that for an instance `i` and an attribute `a`: $\texttt{range}(\texttt{i}, \texttt{a}) = \texttt{required}(\texttt{i}, \texttt{a})$.

Now, we explain how relations are handled. A relation is manipulated like an attribute, and this manipulation may be recursive, i.e. the range of a relation is a class that can be in turn in relation with another class, and so on. For example, suppose that $\mathcal{D}$ = {`P1,P2,P3`} and that $\mathcal{A}$ = {`exposure, profession`}. The objects in $\mathcal{D}$ share the relation `profession` for which they have the set of values $\mathcal{D}'$ = {`Pro1, Pro2,Pro3`}. The objects in $\mathcal{D}'$ share the attribute `exposure`, whose range is `asbestos`. Thus, the class $\texttt{MGI}_{\mathcal{A},\mathcal{D}}$ with the most general intension associated with $\mathcal{A}$ and $\mathcal{D}$ is described by the following constraints: $\texttt{range(profession,MGI}_{\mathcal{A},\mathcal{D}}) = \texttt{MGI}_{\mathcal{A},\mathcal{D}'}$ and $\texttt{range(exposure,MGI}_{\mathcal{A},\mathcal{D}'}) = \{\texttt{asbestos}\}$.

3.2 Domain Knowledge

One originality of the present work is to take into account the knowledge associated with the domain being analysed. Actually, knowledge is encoded into classes and the conceptual hierarchy. Given a conceptual hierarchy $\mathcal{H}$ and a set of objects `E`, it is possible to find the most specific class `C` w.r.t. $\preccurlyeq$ in $\mathcal{H}$ which subsumes all the objects of `E`. According to this remark, given a set of objects $\mathcal{D}$ and a conceptual hierarchy $\mathcal{H}$, we introduce a special attribute `a-kind-of` whose value is the most specific class the objects of $\mathcal{D}$ are subsumed by.

Let us examine how the class hierarchy can be exploited by means of an example. Let us suppose that $\mathcal{D}$ = {`P1,P2,P3`} and $\mathcal{A}$ = {`cancer,profession,prescriptions,exposure`}. The view point $\mathcal{H}_{\mathcal{D},\mathcal{A}}$ relying on $\mathcal{A}$ and $\mathcal{D}$ is shown on the figure 2. The class `C1` is a subclass of `Patient`, as indicated by the special attribute `a-kind-of`, i.e. every instance of `C1` is also an instance of `Patient`. Moreover, the attribute `profession` establishes a relation between the class `C1` and the class `C6`, `C6` being a subclass of `Handwork`.

3.3 Discussion

One of the major advantage of the present work is the ability to handle structured objects. Thus, viewpoints are not only described by boolean attributes (like it is the case in classical approaches) but by attributes with a set of possible (range) and compulsory (required) values. Moreover, the manipulation of relations allows to construct several viewpoints linked by relations. In our application, the OBRS formalism allows us to significantly decrease the number of attributes used to describe children and then increases the efficiency of the viewpoint construction algorithm. Indeed, a child can be affected by twenty different kinds of illnesses and each kind of illness can be attended by at least five treatments. The illnesses are then described by a hierarchy of classes, and each different kind of illness is describe by its possible treatments. A relation is then simply defined between a child and an illness. In the classical data representation, we would have to introduce one attribute for each illness and one attribute for each treatment associated with an illness which represents at least one hundred attributes. Moreover, an illness can be attended by more than one treatment what can not be taken into account in the classical representation without a significant increase of the number of attributes

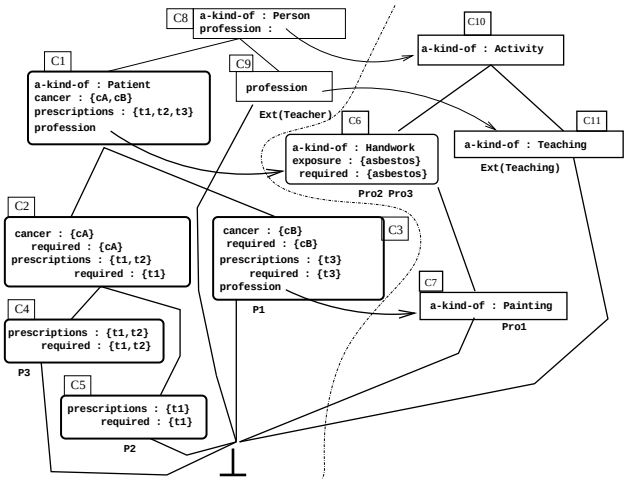

Figure 2: Galois lattice of $\mathcal{H}_{\mathcal{D},\mathcal{A}}$.

(one attribute for two kinds of treatments and so on).

The ability to handle classes to construct viewpoints is a real improvement. Indeed, the Galois lattice technique is time and space expensive and can not be used with large sets of objects (it is a real limitation for KDD purpose). Thus, in order to analyse large databases, data are previously clustered into classes. Classes can be obtained from experts, from reification of requests or from conceptual clustering methods based on statistical measures [7]. For example, in our application, the analyst has defined six classes of different types of childhood cancer which resume all the data (nine hundred children). These classes are then used to generate viewpoints. One can note that this number of data can not be handled by classical method in tractable time and space [6].

A viewpoint is a potential knowledge for the analyst. Indeed, a class is able to represent a concept previously unknown and the hierarchical structure of viewpoints conveys knowledge about the organisation of data. For example, the construction of a viewpoint on the six types of childhood cancer has put in evidence that solid tumours are close to leukemia with respect to one type of treatment, and that three types of solid tumour can be generalised in one type with respect to the kind of treatment used to attend them.

In our approach, the same formalism is used to represent the results of the learning process, background knowledge and data. A viewpoint is then a conceptual hierarchy similar to the one defined in section 2. So, as viewpoints are conceptual hierarchies, they can be stored in the knowledge base, and the classification process of OBRS is used to draw inferences on data and for reasoning purpose [2]. Another knowledge about data is association rules. The following section is about association rule extraction process.

4 Association Rules Extraction

In this section, we show how association rules can be extracted within and between viewpoints to finish with a discussion on the rule extraction process.

4.1 Rules within Viewpoints

An association rule $\mathtt{B}' \rightarrow \mathtt{C}'$, where $\mathtt{B}'$ and $\mathtt{C}'$ are conjunction of properties, means that every individual that verify all the properties of $\mathtt{B}'$ also verify the properties of $\mathtt{C}'$. In [6], it is shown that an association rule $\mathtt{B}' \rightarrow \mathtt{C}'$ exists iff the most general class w.r.t. $\preccurlyeq$, containing `B'` is subsumed by the most general class containing `C'`. Indeed, an association rule $\mathtt{B}' \rightarrow \mathtt{C}'$ can be generated from a class α as soon as `B'` is a subset of $\mathtt{A}(\alpha)$, i.e. the own properties of α, and `C'` is a subset of the *global properties* of α (own and inherited properties).

In our framework, classes of viewpoints are not only represented by properties but also by attributes associated with ranges and constraints. Then, an element of an association rule is not only a property, but also an attribute with which is associated a conjunction of values (required) and a disjunction of values (range). For example, in the figure 2, the class `C2` has two own attributes: the attribute `cancer` and `prescriptions`. The association rule `cancer: cA` $\leftrightarrow$ `prescriptions:(t1)` $\wedge$`(t1` $\vee$ `t2)` is generated. This association rule can be interpreted as follows: the treatment `t1` is always used to attend cancer `cA`. But, treatments `t2` can also be associated with `t1` to attend `cA`.

4.2 Rules between Viewpoints

The generation of association rules involving the `a--kind-of` relation is more classical, but it shows how knowledge organisation can be exploited in the rule generation process. A classical way of interpreting the `a-kind-of` relation is the following: when the class α is `a-kind-of` $\{\beta, \gamma\}$, then every instance of α is an instance of β and γ. As a matter of consequence, this implies that all objects having the properties of α have also those of β and γ. Since an own property represents a necessary condition, it is possible to conclude that objects having the own properties of α have also the own properties of β and γ. So, the right side of

the rule can be reduced to the own properties of the a-kind-of classes. For example, the rule exposure: asbestos $\rightarrow$ own properties of Handwork, can be generated from the class C6 and the property a-kind-of.

Association rules associated with relations between classes can also be defined between viewpoints. When a relation r is defined for objects lying in a set $\mathcal{D}$, then objects in $\mathcal{D}$ are in correspondence with objects in $\mathcal{D}' = \bigcup_{o \in \mathcal{D}} \texttt{range}(\texttt{r}, \texttt{o})$. Moreover, classes in the view point $\mathcal{H}_{\mathcal{D},\mathcal{A}}$ are in correspondence with classes in the view point $\mathcal{H}_{\mathcal{D}',\mathcal{A}}$. In this case, the own properties of α entail the properties of β, where β defines $\texttt{range}(\texttt{r}, \alpha)$ for the relation r in α. For example, the following rule can be extracted from the class C3: cancer: cB $\rightarrow$ range(profession) with exposure: asbestos.

4.3 Discussion

The association rules extracted from viewpoints have a good degree of expressiveness. For example, the association rule generated from class C2: cancer: cA $\leftrightarrow$ prescriptions: t1 $\wedge$ (t1 $\vee$ t2) can be interpreted as follows: when a person has a cancer cA he is attended with the treatment t1 who can be combined with the treatment t2.

Viewpoints can be considered for their own or in association with other Viewpoints. For example, the following rule can be extracted from the relation between class C3 and class C7: a patient ill with cancer cB has an activity implying painting and an exposure to asbestos. Moreover, the relation a-kind-of is used to extract rules between the viewpoint under construction and the already stored viewpoints. One can note that viewpoints and association rules are closely associated with the object-based formalism used here. Thus, if the analyst agrees, viewpoints or association rules can be saved in the knowledge base in order to be reused.

5 Conclusion

In this paper, we have presented an approach to knowledge discovery in databases in the context of OBRS. In the current approach, data are represented by objects which are then clustered into classes; classes are themselves organised into viewpoints giving a particular view of data represented. Viewpoints are also used to generate association rules that can be used to explain the data; viewpoints and association rules generation are basic tool for a KDD process.

The current work extends previous works on Galois lattices in a number of directions: viewpoints as partial lattice, viewpoints are obtained from classes or instances, association rules within and between viewpoints, and exploitation of domain knowledge in the KDD process.

We plane to implement a technique based of the number of instances in the extension of the classes to generate *partial association rules*. B$'$ $\rightarrow$ C$'$ is a partial association rule with confidence c if c% of individuals that verify B$'$ also verify C$'$. First, Galois lattice will provide a theoretical framework for the partial association rules generation. Moreover, we will obtain the advantages of OBRS on the boolean properties which are used in [1].

References

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast Discovery of Association Rules. *Advances in Knowledge Discovery and Data Mining*, pages 307–328, 1996. AAAI Press.

[2] F. Baader, B. Hollunder, B. Nebel, H.-J. Profitlich, and E. Franconi. An Empirical Analysis of Optimization Techniques for Terminological Representation Systems. *Journal of Applied Intelligence*, 4(2):109–132, 1994.

[3] R.J. Brachman and T. Anand. The Process of Knowledge Discovery in Databases. *Advances in Knowledge Discovery and Data Mining*, pages 37–57, 1996. AAAI Press.

[4] C. Carpineto and G. Romano. Galois: An order-theoretic approach to conceptual clustering. In *Proceedings of the 10th International Conference on Machine Learning (ICML'93), Amherst.* Morgan Kaufmann, 1993.

[5] V. Duquenne. On lattices approximations: Syntactic aspects. *Social Networks*, 18:189–199, 1996.

[6] R. Godin and R. Missaoui. An incremental concept formation approach for learning from databases. *Theoretical Computer Science*, 133(2):387–419, 1994.

[7] A. Ketterlin, P. Ganarski, and J.J. Korczak. Conceptual Clustering in Structured Databases: a Practical Approach. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pages 180–185, 1995.

[8] R. Wille. Concept lattices and conceptual knowledge systems. *Computers & Mathematics with Applications*, 23(6–9):493–515, 1992.